Assessing the practising physician using patient surveys: a systematic review of instruments and feedback methods

Richard G Evansa, Adrian Edwardsb, Sean Evansb, Benjamin Elwyn and Glynevc


**Background.** Individual physician performance assessment is a vital part of the medical regulation debate. In this context, the patient perspective is seen as a potentially valid component. Yet, the theoretical and empirical evidence base for such patient assessments is unclear.

**Objectives.** To identify and evaluate instruments designed to assess patients' experiences with an individual practising physician, and to provide performance feedback at the individual level.


**Study selection.**

Inclusion: (i) completed by patients; (ii) assess practising doctors; (iii) have capacity to assess individual doctors for performance feedback; and (iv) used for individual performance feedback.

Exclusion: (i) completed by colleagues, observers or third parties; (ii) assess medical students, nurses or non-physicians; (iii) assess purely at an organizational level; and (iv) not been used for individual feedback.

All electronic outputs were independently assessed by three reviewers. Data were extracted independently by two of three reviewers using a defined template.

**Results.** Six instruments met the inclusion criteria. They all combine evaluation at both organizational and individual level and implementation methods lack standardization. There is limited data on their construct validity or correlations with other attributes. The purpose and method of individual feedback are not well specified, and the evidence to date about the effectiveness of feedback to obtain improvement indicates professional resistance.

**Conclusions.** For formative goals, more clarity is needed about the aim of providing patient assessments feedback to individual doctors: ‘who’ should do it and ‘how’ to do so to best effect. We need to know whether feedback improves doctor performance and how these evaluations correlate with other physician attributes. For summative purposes more research is required on validity and reliability.

**Keywords.** Patient surveys of clinician performance, performance assessment, feedback of patient assessment.

**Introduction**

Physicians unlike the children of Keillor's Lake Wobegon are not all above average. There is indeed concern that substandard practice exists and that physician performance needs to be measured regularly. However, there is little, if any, consensus on how the performance of practising physicians should be...
measured and by whom—by other doctors, by patients or by independent third parties. Our recent review concluded that the existing methods of assessment by peers lack proven validity, although there is a recognition that doctors ‘know’ the identity of poor performers. Nevertheless, formalizing the assessment of physicians’ performance has assumed some urgency with a focus on how we should make valid and reliable judgements. As we found, the validity of peer-based judgements is open to question. What then is the role of patient-based assessments, given that they experience the performance of doctors at first hand? In the UK, a contractual incentive stimulated general practitioners to obtain data using recently developed and recommended patient survey instruments, while in the US the Maintenance of Certification (MOC) program is encouraging use of confidential patient satisfaction measures as one element of assessing performance in practice. As the regulation debate continues in both countries and is mirrored internationally, it is important to assess the science underpinning these instruments in more detail.

The patient view began to emerge as a potential means of feedback for service improvement from the 1980s drawing on consumer survey literature from the previous 20 years. Yet, it is only in the last 5 years that the potential for educational feedback at the level of an individual practitioner has come into focus. The need for valid and reliable performance feedback to assess fitness to practice adds another focus to this patient source. The increasing importance of the patient view as part of a ‘360-degree assessment’ warrants a review of available instruments and their potential contribution to performance assessment.

We sought to identify and evaluate existing instruments for patient assessment of their experience with an individual doctor in routine practice that have been used for feedback to individual doctors, to evaluate their development, validity and reliability, as well as their potential to identify poor performance of the individual doctor for further assessment.

Methods

Search strategy

A systematic search was performed for published material using Medical Subject Headings and text words in the following nine electronic databases: PubMed (1985-), Embase (1985-), PsycInfo (1985-), SIGLE (1985-), HMIC (1985-), ASSIA (1985-), CINAHL (1985-), Cochrane (1985-) and Dare (1985-). Initial searches identified little work on patient surveys at the level of individual episodes of care prior to 1985 where the focus was at the level of organization or service. The search strategy used for all the databases used combinations of text words and MeSH terms found to be relevant from pilot searches; there were no language restrictions. The text word combinations used were the following: ‘patient satisfaction’ or ‘patient participation’ or ‘physician-patient relations’; ‘physician’ or ‘doctor’; and ‘questionnaire’ or ‘survey’ or ‘instrument’ or ‘feedback’ or ‘evaluation’. For PubMed and Embase a further combination of ‘Professional competence’ was added as text and MeSH term to narrow the search for these databases. Reference lists from review articles and included papers were searched for additional instruments. The initial search was run in July 2004 and rerun November 2005.

Inclusion and exclusion criteria

To be included, the instruments needed to be published in the peer-reviewed literature and fulfil the following criteria:

- (i) be completed by patients;
- (ii) assess practising doctors;
- (iii) have capacity to assess individual doctors for performance feedback; and
- (iv) have been used for individual performance feedback.

Instruments were excluded if any of the following were evident:

- (i) completed by colleagues, observers or third parties;
- (ii) they assessed medical students, nurses or non-physicians;
- (iii) they assessed purely at an organizational level; and
- (iv) they had not been used for individual feedback.

Selection

Three reviewers (RGE, SE and BE) assessed all electronic outputs (abstracts) independently. Full-text articles considered as possibly relevant were assessed by three reviewers (RGE, GE and AE). Selection against the criteria was then decided by the reviewers together. Where it was unclear from the literature whether the instrument had the capacity to assess an individual, the authors were contacted for further information. The decision on final inclusion against criterion 4 was made on the basis of the published data backed up by email contact with all the authors to identify any missed data from the peer-reviewed literature. This included asking the authors to check whether we had omitted any instruments from the list.

Data extraction and synthesis

A template was piloted and amended for data extraction. The template covered the following areas: conceptual basis for the instrument’s development, the measurement construct, piloting and validation.
process, characteristics of patient and physician sample, psychometric data, setting of main use and description of feedback process. One researcher (RGE) applied the template to all included studies; a second researcher (GE or AE) also undertook data extraction for each instrument. The data were summarized and a synthesis conducted by RGE, GE and AE.24

Results

Figure 1 summarizes the results of the search.

We identified 3476 studies from the search strategy on nine databases (SIGLE, HMIC, PsycInfo as one database bundle 3937, ASSIA 831, CINAHL 649, Cochrane 265, Dare 1367, Embase 245 and PubMed 1432) after the two search outputs were combined and duplicates removed.

In summary, 119 papers were retrieved for full-text assessment; 55 were excluded by a single reviewer (RGE) with no instrument being described. From the remaining 64 papers we identified 25 potential instruments. Citation follow-up identified a further nine instruments, making a total of 34 instruments for detailed assessment. Eleven instruments were excluded because they did not have an explicit aim to provide data for individual physician feedback. 1 lacked psychometric data and 16 were excluded because there was no evidence that they had been used to provide individual feedback. Six instruments met all inclusion criteria.

Instruments excluded

No focus on individual performance. The 11 instruments lacking an explicit aim to provide data for individual physician feedback are listed as follows: Primary Care Assessment Survey,25 Practice Accreditation and Improvement Survey,26 Consumer Assessment of Health Providers and Systems,27,28 Components of Primary Care Index,29 Quality from Patients Perspective,30 Patient Satisfaction Questionnaire,31 Parents Perceptions of Primary Care Measure,32 and questionnaires by Bitzer et al.,33 Gericke et al.,34 Bredart et al.,35 and Etter and Perneger.36 For the Visit-Specific Questionnaire and Patient Satisfaction Index37 psychometric data were unobtainable and excluded at this stage, making 12 excluded in total.

Not used for individual performance feedback. The 16 instruments excluded because there was no evidence as yet that they had been used for providing individual feedback are as follows: Consultation and Relational Empathy Measure (CARE),38,39 Patient Enablement Instrument (PEI),40–44 Consultation Quality Index (CQI),41 Consultation Quality Index-2 (CQI-2),35 Patient-Doctor Relationship Questionnaire,46 Medical Interview Satisfaction Scale,47,48 Patient Experience Questionnaire,49,50 Medical Outcomes Study Questionnaire,51 Patient Intentions Questionnaire/Expectations Met Questionnaire,52 Physicians’ Humanistic Behaviours Questionnaire,53 Patient-Doctor Interaction Scale54 and unnamed instruments developed by authors Crossley et al.,55 Haddad et al.,56 Wooliscroft and Howell,57 Matthews and Feinstein16 and Feletti et al.15 Many of these 16 excluded instruments do not appear to have been developed beyond their initial research findings, perhaps highlighting a lack of interest in their implementation and associated costs. The PEI40–44 evaluates the doctor as being more or less ‘enabling’ in terms of the patient’s well-being and coping status and has also been combined with data on length of consultation and how well the patient knows the doctor in the CQI.41 However, it is essentially a patient-based outcome measure58 rather than an evaluation of the patient’s experience of visiting the doctor and does not have specific data regarding experience of its use in feedback to individual doctors. The CARE instrument developed by Mercer38,39 aims to give individual physicians direct feedback on their relational empathy as perceived by the patient and is being piloted as part of a planned new licensing examination in the UK.59 The CQI and CARE have now been combined as the CQI-2.45

Instruments included

Six instruments with clear description of development and validation and which have been used for physician feedback are included37,60–63; these are the Consultation Satisfaction Questionnaire (CSQ),17,64,65 the Chronically Ill Patients Evaluate General Practice (CEP),60 the General Practice Assessment Questionnaire (GPAS, now GPAQ),62 the Patients Evaluate General/Family Practice (EUROPEP),61 Improving Practice
CSQ was developed by Baker in 1990.\textsuperscript{17,64,68,70,79} It is a Consultation Satisfaction Questionnaire. The 18-item CSQ was developed by Baker in 1990.\textsuperscript{17,64,68,70,79} It is a quantitative assessment of patient satisfaction from a specific visit with a specific doctor in relation to four domains: general satisfaction, professional care, depth of relationship and perceived time. A separate tool was developed for evaluating the organizational aspects, namely the Surgery Satisfaction Questionnaire.

**Consultation Satisfaction Questionnaire.** The 18-item CSQ was developed by Baker in 1990.\textsuperscript{17,64,68,70,79} It is a quantitative assessment of patient satisfaction from a specific visit with a specific doctor in relation to four domains: general satisfaction, professional care, depth of relationship and perceived time. A separate tool was developed for evaluating the organizational aspects, namely the Surgery Satisfaction Questionnaire.

**Chronically Ill Patients Evaluate General Practice.** The CEP, developed by Wensing et al.,\textsuperscript{19,60} is a 51-item questionnaire for evaluating 10 dimensions of care; it has a mix of organizational- and practitioner-level evaluation items (appointments/emergencies, premises, continuity, cooperation, medical care, competency/accuracy, relation and communication, privacy, information and advice and support). Providing individual feedback to doctors did not lead to any improvement on the doctors’ communication behaviour, and many doctors perceived the patient views as less relevant than before the study.\textsuperscript{19} A very modest correlation was found between the CEP and IPQ,\textsuperscript{19,77,78} suggesting a need for more work on construct validity.

**General Practice Assessment Questionnaire.** The GPAS, now GPAQ, is a 53-item questionnaire for evaluating five areas of care: access, technical competence, effective GP communication, GP interpersonal abilities and effective organization of care.\textsuperscript{62,73,80,81} It has a mix of organizational and ‘usual doctor’ evaluation which is not necessarily visit specific. The shorter 36-item GPAQ has a postal version which is not visit specific and a post-consultation version which is visit specific.

**EUROPEP.** The EUROPEP instrument is a 23-item questionnaire for evaluating five dimensions of care: relation and communication, medical care, information and support, continuity and cooperation, access and facilities.\textsuperscript{61,72,75} It has a mix of organizational and ‘usual’ practitioner evaluation that is not visit specific but is to be considered ‘over the last 12 months’.

**Improving Practice Questionnaire.** The IPQ is a 27-item questionnaire combining evaluation of two components of care: the Doctor’s Interpersonal Skills Questionnaire (DISQ) and service issues of access and environment.\textsuperscript{18,20,26,63,76,82–85} It has a mix of organizational and individual visit-specific evaluations.

**Provider-Specific Survey.** The PSS is a 34-item scale combining a visit-specific evaluation of five aspects of care: making the appointment, arrival at the office and registration, waiting to see the clinician, service by staff other than the main clinician and seeing the clinician.\textsuperscript{66,67}

**Synthesis**

Although there are several instruments that use patient evaluations, the aim of the individual feedback from these instruments is difficult to discern beyond a general intention to measure ‘improvement’ along the scales of the items being rated. There are none that clearly state an aim to use patient assessments of individual doctors to provide direct summative judgments against a standard and there is usually an implied use for formative or developmental performance improvement. Those that actually provide individual assessments have only recently been used in this way, and data on evaluating the effect of feedback were limited to two instruments CEP and IPQ.\textsuperscript{19,77,78} How the feedback process is conducted, by whom, to whom and in what setting are not specified. Methods of feedback to individuals are unspecified, beyond reflecting on the normative data, with or without the help of a colleague. The implicit assumption seems to be that normative comparisons will motivate change in professionals because it is believed they are sensitive to patient views about their performance. However, clinician reaction to the feedback of patients’ assessments has not been positive.\textsuperscript{19} Although there are less than a handful of studies on this issue, there is evidence that clinicians can be defensive and resistant, unless feedback is provided during early training grades.\textsuperscript{18,20} These limited data suggest that the setting and feedback process are crucial to achieve any effective change for doctors in practice.

Five of six included instruments have been developed and evaluated in European primary care and one US instrument in office-based primary and specialist care. A common theme with all instruments, except the CSQ,\textsuperscript{17} is their attempt to evaluate at both the organizational and individual level in a single instrument. While explicit labelling of items helps differentiate these aspects, there is scope for crossover bias and potential for patients to be unclear about whether they are primarily evaluating the organization or an individual clinician. Another source of potential confusion is the variation between different methods of implementation such as immediate post-consultation use and postal survey with consequent variations in score. Sometimes more than one implementation method is used for the same instrument. Different implementation gives rise to different evaluations, which will undermine benchmarking and comparative feedback.

There are very little data on how patient scores correlate with other assessments of a practising physician, suggesting a need for more work on construct validity. A very modest correlation was found between the
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Measurement aim or construct</th>
<th>Scale design (no. of items)</th>
<th>Implementation details</th>
<th>Instructions to rater</th>
<th>Manual</th>
<th>Commercial or free</th>
<th>Published benchmarks</th>
<th>Required sample size for stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSQ, Baker, 1991</td>
<td>Empirical assessment of patient satisfaction</td>
<td>Five-point Likert agreement scale (18)</td>
<td>Given on arrival, completed post-consultation before leaving.</td>
<td>No data</td>
<td>No</td>
<td>Free</td>
<td>Feedback given with anonymized scores for other participants. Published aggregate means.</td>
<td>Given to 70 patients for 40 returns—reliability 0.91 from 40 patients.</td>
</tr>
<tr>
<td>CEP, Wensing, 1998</td>
<td>Patient satisfaction with 11 dimensions of care: appointments, emergency availability, premises, continuity, cooperation, medical care, competence/accuracy, relation and communication, privacy, information and advice, support</td>
<td>Six-point Likert, poor to very good (51)</td>
<td>Two methods: handing out to attenders or mailing to registered patients—no detailed data.</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td>Mean figures per dimension plus lowest and highest in range.</td>
<td>No data. Feedback report based on 100 consecutive patients per doctor.</td>
</tr>
<tr>
<td>GPAS—now GPAQ—Ramsay et al., 2000</td>
<td>Five key areas: (i) access; (ii) technical competence; (iii) effective GP communication; (iv) GP inter-personal abilities; and (v) effective organization of care</td>
<td>Six-point Likert, very poor to excellent (GPAS 53, GPAQ 36)</td>
<td>GPAS used pre-consultation and as postal survey. GPAQ to be used post-consultation at exit or as postal survey.</td>
<td>Written instruction at head of questionnaire</td>
<td>Manual—downloadable from website</td>
<td>Free—website includes software for analysis of results.</td>
<td>National benchmarks on website—based on postal survey results of GPAS—being updated by data from GPAQ use.</td>
<td>55. GPAQ website recommends 50 for individual feedback.</td>
</tr>
<tr>
<td>EUROPEP instrument—Grol, 2000</td>
<td>Five dimensions of care—relation and communication, medical care, information and support, continuity and cooperation, facilities availability and accessibility</td>
<td>Five-point Likert, poor to excellent (23)</td>
<td>Variable—postal return or waiting room completion.</td>
<td>Written instruction page</td>
<td>No</td>
<td>Free—request to register with EUROPEP and contribute to database</td>
<td>Reference figures available through registering.</td>
<td>No specific data. Aim for sample of about 60 patients.</td>
</tr>
<tr>
<td>IPQ—Greco, 2003</td>
<td>Patient satisfaction with two components: (i) DISQ; and (ii) service issues—access, environment, staff information</td>
<td>Five-point Likert, poor to excellent (27)</td>
<td>Post-consultation before leaving.</td>
<td>Written instruction at head of questionnaire.</td>
<td>Published handbook—‘IPQ Toolkit’</td>
<td>Commercial via CFEP-UK surveys</td>
<td>Published means with 95% confidence intervals.</td>
<td>30–50 suggested—no published data.</td>
</tr>
<tr>
<td>PSS—Nelson, 2004</td>
<td>Patient satisfaction with two dimensions of care. (i) clinician care; and (ii) features of visiting the office (making the appointment, arrival and registration, waiting, service by staff other than clinician)</td>
<td>Five-point Likert, poor to excellent (34)</td>
<td>Mailed to randomly selected patients following appointment with doctor.</td>
<td>Covering letter</td>
<td>No</td>
<td>No data</td>
<td>Reliability of 0.80 at clinician level (11-stem clinician care scale) from 66 patients.</td>
<td></td>
</tr>
<tr>
<td>Conceptual framework</td>
<td>Literature review to identify relevant patient issues, discussion with GPs and list of statements agreed. Six iterative versions developed with colleagues, response patterns studies, editing process guided by principal components analysis.</td>
<td>Correlation coefficient with Medical Interview Satisfaction Scale 21, 0.82. Significant predictive construct validity (patients changing practice). Principal components analysis, final version, revealed three factors: (i) professional care; (ii) depth of relationship; and (iii) length of consultation.</td>
<td>Cronbach’s alpha, 0.91. Test–retest data in 131 patients, Pearson product moment correlation, 0.82. Total assessments: 9717 (6 studies). (response rate, 76.6%)</td>
<td>No detailed data but response rate indicates reasonable feasibility</td>
<td>Adult patients in primary care—consecutive consultations. Mean age 45.5 years. Female 66.4%</td>
<td>126 GPs in 39 practices in south-west region of England. The sample contained 39 trainers in 84 training practices.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSQ, Baker, 1991</td>
<td>No clear link identified to social or psychological theory.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEP, Wensing, 1998</td>
<td>Dimensions derived from systematic literature review, 57 studies, plus a qualitative focus groups study of GPs and of patients with chronic illness.</td>
<td>No data on reliability indices; 762 assessments, (response rate 67%).</td>
<td>No detailed data but response rate indicates reasonable feasibility</td>
<td>Sample of patients from primary care in Dutch practices with chronic disease (COPD, asthma, diabetes, CVD, etc.). Demographic data on sample not available.</td>
<td>Sample of patients from primary care in Dutch practices with chronic disease (COPD, asthma, diabetes, CVD, etc.). Demographic data on sample not available.</td>
<td>28 GPs in Netherlands. No further data.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPAS (now GPAQ), Ramsay, 2000</td>
<td>Patient-derived service delivery priorities, based on work by other authors.</td>
<td>No data. Based on US instrument, Primary Care Assessment Survey.</td>
<td>No data on construct validity or factor analysis.</td>
<td>No detailed data but response rate indicates reasonable feasibility</td>
<td>Sample of patients from primary care in Dutch practices with chronic disease (COPD, asthma, diabetes, CVD, etc.). Demographic data on sample not available.</td>
<td>28 GPs in Netherlands. No further data.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUROPEP instrument, Grol, 2000</td>
<td>List of patient priorities derived from systematic literature review. Items ranked using patient survey to reflect priorities and dimensions of care.</td>
<td>Construct validity not specified directly. Scree plot indicated three components: access, patient-centredness and nursing. Principal axis factor analysis indicated these accounted for 64.1% of variance.</td>
<td>Cronbach’s alpha; communication items 0.90; inter-personal items 0.93; others range 0.69–0.95. Text–retest correlations 0.810.92. 7247 assessments, (response rate 66%).</td>
<td>Survey completion appears possible by patients within estimated 10–15 minutes waiting time67.</td>
<td>Sample of patients from primary care in Dutch practices with chronic disease (COPD, asthma, diabetes, CVD, etc.). Demographic data on sample not available.</td>
<td>28 GPs in Netherlands. No further data.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument</td>
<td>Description</td>
<td>Validity/Reliability</td>
<td>Sample Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPO, Greco</td>
<td>Not described in detail. 'Facilitating patient-centred change' in GP setting. DISQ items derived from literature review and areas identified as important by patient and GP focus groups. Further detail referred to as 'unpublished data'.</td>
<td>No data on construct validity. Predictive validity suggested by higher ratings by patients over 40 years old. DISQ: correlation with Falvo-Smith Interaction Scale, Pearson product moment $r = 0.77$. Principal components analysis: two factors accounting for 69% of variance: (i) clinician’s interpersonal skills; and (ii) access, staff attitudes and information. DISQ: one factor accounted for 70.31% of variance.</td>
<td>Completion time average approximately 3 minutes. Single A4 sheet.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSS—Nelson</td>
<td>No data on construct validity. Scree test and Tucker’s rho suggested two underlying dimensions. Total variance accounted for by two factors, 91% primary care, 92% specialty care.</td>
<td>10–15 minutes</td>
<td>Mean age 51, 58% in age range 21–60 years. Females, 64.8% and 73.9% visiting usual GP. Approximately 23% less than 5 years duration of experience of the practice.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:**
- IPO: IPQ, Greco, 2003 (including DISQ, Greco 1999)
- DISQ: correlation with Falvo-Smith Interaction Scale, Pearson product moment $r = 0.77$. Principal components analysis: two factors accounting for 69% of variance: (i) clinician’s interpersonal skills; and (ii) access, staff attitudes and information. DISQ: one factor accounted for 70.31% of variance.
- PSS: PSS—Nelson, 2004
- Systematic review of instruments and feedback methods.
DISQ scores and expert assessment of training grade doctors, while no correlation was found between the CSQ and performance in video assessment in the Membership of Royal College of General Practitioners exam. This clearly needs further research if data were to be used for performance management or assessment of fitness to practice.

Discussion

Six validated instruments met the criteria for patient assessment of their experience with an individual physician and which have been used for feedback to individual physicians. There is no clear statement of purpose for each instrument as a formative or summative assessment. All were developed in primary care or outpatient setting and appear to have the potential for identifying poorly performing doctors. However, there are as yet few data on which this could be based with confidence and only the PSS explicitly attempts to address dealing with low scores. Other than the early CSQ, all the instruments have varying methods of implementation and include evaluations of the organization or service alongside individual evaluation, leading to difficulty in interpreting their findings.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Aim of feedback</th>
<th>Methods of feedback</th>
<th>Evaluation of effect of feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSQ, Baker, 1991</td>
<td>'To help improve consulting skills' by promoting responses to feedback such as distance learning or small group work</td>
<td>Feedback of individual doctor scores provided, with anonymous comparison scores. No further details.</td>
<td>No data</td>
</tr>
<tr>
<td>CEP, Wensing et al., 1998</td>
<td>‘Feedback was expected to result in the undertaking of activities for improvement, e.g. communication training.’</td>
<td>Report included a summary of practice results and article suggests that the data were not individualized: ‘no GP specific interpretation of results or GP specific advice’. Authors suggest feedback was a peer review group activity.</td>
<td>Patient evaluations of individual doctors using CEP did not lead to improvement—despite range of reported actions in response to feedback. Evidence of increased resistance and defensiveness among doctors.</td>
</tr>
<tr>
<td>GPAS (now GPAQ), Ramsay et al., 2000</td>
<td>To identify ‘some behaviour which you could think about improving’, e.g. doctors’ communication skills in consulting.</td>
<td>Website software available for analysis at practice and individual level with normative benchmarks. Suggestions for guided reflection of individual and referred to consultation analysis methods. Suggestions on discussing with colleagues or mentors.</td>
<td>No data.</td>
</tr>
<tr>
<td>EUROPEP, Grol and Wensing, 2000</td>
<td>To ‘help identify opportunities for improvement’. Overall aim to ‘assess quality and provide relevant feedback, at different levels’.</td>
<td>Different types of feedback reports developed by EUROPEP participants: Swiss, German, Portuguese examples with benchmarking. Authors discuss need for further research.</td>
<td>No data.</td>
</tr>
<tr>
<td>IPQ, Greco et al., 2003</td>
<td>To ‘be used as a meaningful source for facilitating patient-centred change’ “to highlight areas in need of improvement across British general practice,” To identify ‘strengths surrounding their interpersonal skills and areas which could be targeted for personal development’.</td>
<td>Client-Focused Evaluation Program—UK surveys provide IPQ feedback at practice and at individual level with normative benchmarks. DISQ element of IPQ originally designed for evaluating performance and providing individual feedback in an educational setting. Suggestions for consultation analysis methods.</td>
<td>Repeat survey after detailed feedback: no significant change. Attitudes of physicians to feedback varied. Two studies evaluating effect of DISQ feedback on GP registrars. Both found significant improvement in intervention groups. Regular systematic feedback effective more in earlier stage of training.</td>
</tr>
<tr>
<td>PSS—Nelson et al., 2004</td>
<td>To “help individual learners better understand and improve their relationships with patients.” Feedback allows comparison with other clinicians in section or department, and across system. It is also used for both clinician self-assessment and as a component of the annual performance review.</td>
<td>“Individual clinicians receive graphic and tabular satisfaction feedback reports 2 time per year.” Consistent low scorers are advised to receive personal coaching or attend national seminars related to patient interaction.</td>
<td>No data</td>
</tr>
</tbody>
</table>
The purpose for assessing must be clear. To be used in a summative way for decision making, the science underpinning the instruments needs to be more robust. The construct validity of these instruments in terms of correlation with other assessment perspectives needs further research. The lack of any meaningful correlation between one form of assessment (the video component of the MRCGP) and patient feedback via the CSQ or PEI highlights this.86 If the purpose is formative, then we need a better scientific understanding of the effect of feedback and how to best use it.

The evidence of the effect of feedback on practising physicians is limited to two of the five instruments19,77,78,87 where no improvement was shown and a mixed reaction from doctors, one study generating negative perceptions in the feedback group.19,87 The only evidence of positive effect relates to the DISQ element of the IPQ when used in a structured educational setting with GP registrars early in their training with less effect seen later in training.18,20 The insufficient evidence does not allow more comment than that the effect of feedback can be equivocal or negative.

**Strengths and weakness**

We used several databases with delineated search strategies and complementary methods to identify relevant material. However, the search strategy was narrower for the PubMed and Embase databases and non-English language sources were not specifically sought, each potentially limiting recall. Multiple assessments were undertaken at key decision points and dual independent data extraction methods were applied. We focussed our research question on a current issue for the debate around the assessment of practising doctors and we are unaware of any similar analysis.

**Current context**

Quality improvement initiatives classically intervene at the system level to enhance quality and safety in health care. Continuing professional development and appraisal also examine how the individual clinician is performing. In this latter sphere, increasing importance is given to patients’ views of their doctors. In UK primary care, for example these assessments have already been integrated into the incentivized ‘Quality and Outcomes Framework’ of doctors’ contract with the National Health Service5 and form part of the UK Chief Medical Officers policy proposals,14 while in the US the MOC program8,9,88 is being developed on a continuous cycle. In this way, patient evaluations are now becoming part of the landscape of clinical care development—the 360-degree assessment22 that includes patient, peer and other staff evaluations; clinical or health outcome indicators and identifying cost-effective and evidence-based practice, all focussed on the individual clinician.

**Implications**

We suggest that greater clarity is needed about the purposes to which the available instruments may be put. The overall aim of these instruments seems to be formative in identifying the areas of consulting behaviour from patient experience, where there is scope to improve in relation to the patient-centred model of consulting.89 They rely on normative comparisons to motivate change in professionals who are assumed to be sensitive to patient views on their behaviour. Relying on this is questionable as evidence from wider organizational research suggests that such normative comparisons can lead to performance decline as well as improvement.90 There are variable methods for administration and sampling, and policies on data ownership and benchmarking are uncertain. Individual-level feedback methods are also not specified beyond reflecting on the normative data (with or without a colleague), whereas potential responses to feedback in terms of what doctors can do to improve communication skills are described copiously. There is a gap between data for feedback and action to improve, where the precise mechanism of formative individual feedback for clinicians is unclear. Research is particularly required on how to undertake effective feedback processes.

Alternatively, the goal may be summative. This would require a major advance in the validation of the instruments, with further assessment of construct validity in current settings (usually primary care), validation in other settings such as secondary care and analysis of the determinants of variation. In this context, the aim would be to look for outliers who would then be subjected to further assessment of competence or in the extreme scenario, their fitness to practise. Holders of benchmark data would need to consider their response to outliers, and whether they are obliged to disclose about possible poor performers. Responsiveness would also need validating in this high-stakes context where detectable and demonstrable improvement needs to be reliably measured.

**Conclusions**

Patients are the end-users of health care. Clinicians have been lukewarm towards feedback from users,19,78 but this cannot be dismissed. We must continue to measure patients’ assessments of their experience with individual clinicians and try to understand more what they mean, how they correlate with other aspects of individual performance, and how doctors can learn and improve by the assessment. If, however, these assessments become more summative, part of ‘performance management’, and integral to re-certification, then in terms of procedural justice91 there has to be more science around validity, reliability, standardized administration of the instruments and sampling and investment in comparative benchmarking. We also need
to know more about how to use these assessments, who should provide the feedback and which process works best. The available instruments offer a platform from which to develop patients’ assessments of physician performance, as patients are best placed to make judgements about many crucial areas of practice.

Declaration
Funding: None.
Ethical approval: None.
Conflicts of interest: None.

References